Adapting Digital Libraries to Continual Evolution

Bruce R. Barkstrom, Melinda Finch, Michelle Ferebee, and Calvin Mackey
Atmospheric Sciences Data Center, NASA Langley Research Center
Mail Stop 157D, Hampton, VA 23681-2199 USA
+1 757 864 5676
b.r.barkstrom@larc.nasa.gov

ABSTRACT

In this paper, we describe five investment streams (data storage infrastructure, knowledge management, data production control, data transport and security, and personnel skill mix) that need to be balanced against short-term operating demands in order to maximize the probability of long-term viability of a digital library. Because of the rapid pace of information technology change, a digital library cannot be a static institution. Rather, it has to become a flexible organization adapted to continuous evolution of its infrastructure.

Categories & Subject Descriptors: K.6 MANAGEMENT OF COMPUTING AND INFORMATION SYSTEMS

General Terms: Economics

Keywords: Investment streams, data archival, knowledge management, data production control, data transport and security, personnel skill mix, IT evolution

SWIMMING IN A RIVER OF PERPETUAL CHANGE

In an earlier and more tranquil time, a library might be regarded as a quiet haven from the hustle and bustle of the surrounding world. The contents of such a library would be used for meditation and study, with only occasional intrusions from the outside world [3].

In our current world, this isolation is hardly possible. The contents of our digital libraries are used to guide government policy and commercial affairs. The mechanisms for acquiring, holding, and distributing digital library holdings are subject to limited lifetimes and massive change. They are driven by commercial needs rather than those of the library community. We are clearly faced with the Biblical command: "Do not store up for yourselves treasures on earth, where moth and rust consume and where thieves break in and steal;" [4, Matt. 6:19 – one of the early records of the need for information security].

It follows that we cannot view digital libraries and their holdings as institutions with a large initial budget and a smaller maintenance budget. This might be an appropriate view if the assets resembled those of a cathedral or a fort – once construction finishes, the building is impervious to most forces of change and needs only a minimal maintenance budget thereafter. Rather, it becomes critical to recognize the need for continual investment – and to develop the budgetary discipline necessary to preserve the investment streams that allow the library to evolve.

In the material that follows, we identify five critical investment streams and suggest a strategic approach to preserving them.

KEY INVESTMENT STREAMS

In developing a budget strategy for the NASA Langley Atmospheric Sciences Data Center (ASDC), we have been led to the idea that there are five critical investment streams for us to guide:

Data Storage Infrastructure – the disks and robotic tape storage devices that hold our data, as well as the software that manages that infrastructure

Knowledge Management – the data structures (including databases) and documentation that hold the inventory, catalog, metadata, and other search mechanisms, as well as the user interfaces that allow visibility and access to our data

Data Production Control – the software and workflow procedures that allow us to produce data within our data center

Data Transport and Security – the network infrastructure and protective mechanisms that allow us to safely ingest data, move it from one device to another, and that deliver that data to our users

Personnel Skill Mix – the human inventory portion of our data center without which we could not provide our users with the data access they need

In the subsections that follow, we provide more detail on these streams. We are particularly interested in identifying strategies for smoothing the investment flow into each one.

Data Storage Infrastructure

If the data contained in a digital library are its primary asset, then the evolution of the data storage infrastructure is one of the primary driving forces for system evolution. As our system is configured now, the primary data holdings reside in massive, robotic tape storage silos. These storage devices are not well suited to either high bandwidth data streaming or to bursty access patterns. To accommodate these loads, our current system needs to move data from the tape archive onto disks, from which it can serve either the computers that transform our input data streams into more finished scientific data products or to distribute data to our user communities. This evolution clearly will be driven by

JCDL'02, July 24-28, 2002, Portland, OR, USA.
ACM 1-58113-345-6/01/0006.

the continuing improvement in the price-performance curves of disks with respect to robotic tape archives.

Knowledge Management

It is also clear that we face continuing evolution of the ideas regarding how to search for data, of the software appropriate for managing these metadata, and of the user interfaces to the metadata. In addition, we also face the possibility that metadata cannot hold all of the possible information we need, so that content-based searching of huge stores of binary data will be necessary. We already have a situation in which some of NASA's Earth Observing System (EOS) metadata fields are inappropriate to help data users search for some of the files in our data center. To improve the accessibility of our data, we will have to evolve the metadata and its underlying knowledge management.

Data Production Control

Most digital libraries do not generate their own data. If they did, text libraries might serve as sponsors of resident authors. However, the NASA EOS data centers do produce new kinds of data. This has the advantage that the production process can directly populate the metadata stores and more directly contribute to maintaining data provenance. At the same time, there is some increased activity at the data center that is required to deal with the data production, and its quality control.

Data Transport and Security

In the early days of our data center, delivering data on media was an accepted part of our activities. However, it is clear that users prefer to receive data electronically. It is cheaper for us to deliver data in this way, as well. Unfortunately, for a small number of data users, the amount of data they need is so large that network delivery is unlikely in the near future. Accordingly, we need to distribute very large amounts of data with technology that is driven to produce the lowest cost for both the data center and the users. As technology changes, our data delivery mechanisms must keep pace.

Personnel Skill Mix

Finally, we must recognize the importance of providing for the evolution of the data center's personnel. This is perhaps the most important force of those we have listed. We must continually invest resources in keeping our personnel proficient in the technologies that best serve our customers.

THE NEED FOR DISCIPLINE

The hard part about dealing with continual evolution is the necessity for budgetary discipline. It is our experience that operational needs tend to expand to fill the available budget. Thus, the most important strategy we can identify for dealing with investments that need continual evolution is to give first priority to the investments and to then adjust operational budgets to what is left after investing. The second most important strategy is to develop systematic ways of identifying the largest contributors to expense and uncertainty. This information then allows a more quantitative approach to reducing the continuing cost of the digital library [2] and may make it possible to induce cooperative behavior between different institutions [1].

ACKNOWLEDGMENTS

We thank NASA Langley Research Center and NASA's Earth Science Enterprise for their support of the Langley Atmospheric Sciences Data Center and its activities.

REFERENCES

1. Cooper, B. and Garcia-Molina, H. Creating Trading Networks of Digital Archives, in *Proceedings of JCDL'01* (Roanoke VA, June 2001), ACM Press, 353-362.
2. Crespo, A. and Garcia-Molina, H. Cost-Driven Design for Archival Repositories, in *Proceedings of JCDL'01* (Roanoke VA, June 2001), ACM Press, 363-372.
3. Petroski, H. *The Book on the Bookshelf.* Alfred A. Knopf, New York NY, 1999.
4. *The New Oxford Annotated Bible*, **3rd ed.**, M. D. Coogan, Ed., Oxford University Press, Oxford, UK, 2001.